Mathematics Performance in Secondary Education

Yao-Chih Hsu, Xuan-Chun Wang, and Wen-Lee Sin

June 5, 2025

Outline

Introduction

- Dataset: *Math-Students Performance Data* from Kaggle (Shamim, 2025).

Introduction

- Dataset: *Math-Students Performance Data* from Kaggle (Shamim, 2025).
- Variables `G1`, `G2`, `G3`, and `absences` were provided by the school.
- Remaining variables were collected via questionnaires and are mostly categorical.

Goal

Analyze `G3` to identify influential variables.

Data Grouping

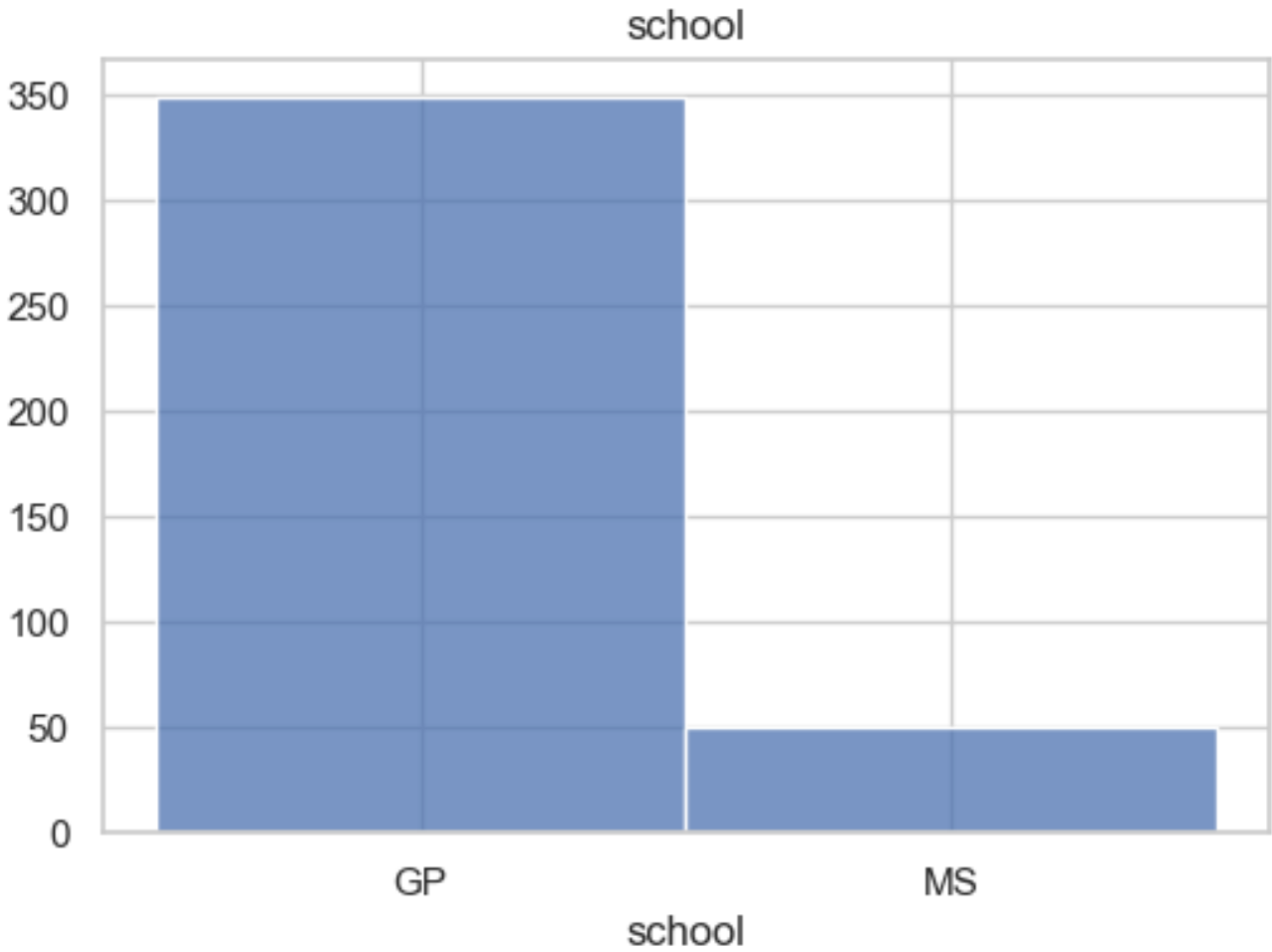
We grouped the data into the following:

Groups	**Variables**
support	schoolsup, famsup, paid
family	address, famsize, Pstatus, guardian, traveltime, famrel
parents	Medu, Fedu, Mjob, Fjob
performance	failures, studytime, absences
alcohol	Dalc, Walc, health
after_class	activities, freetime, goout
school_choice	reason, nursery, higher
score	G1, G2, G3

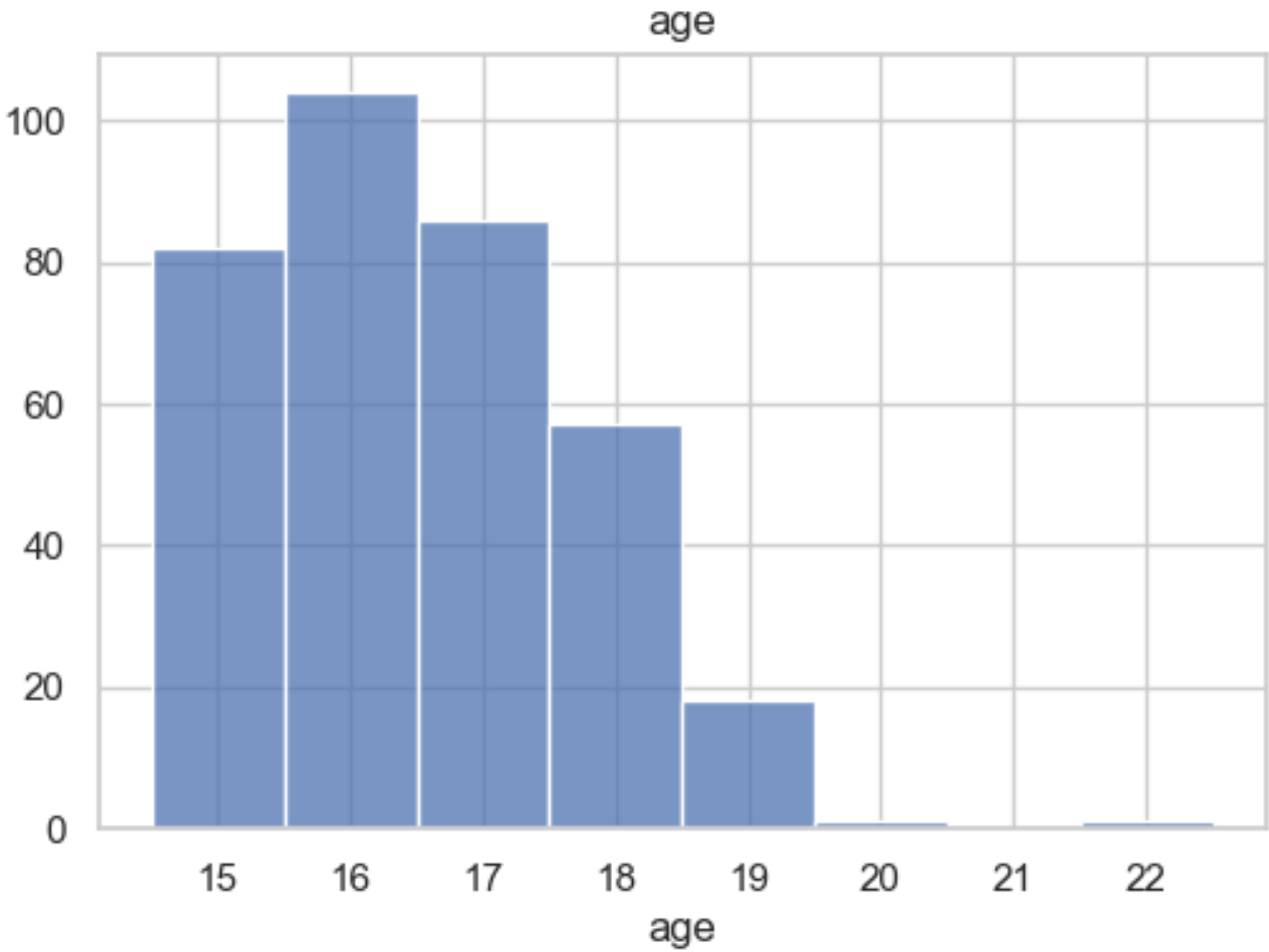
The variables not yet assigned to any group are:

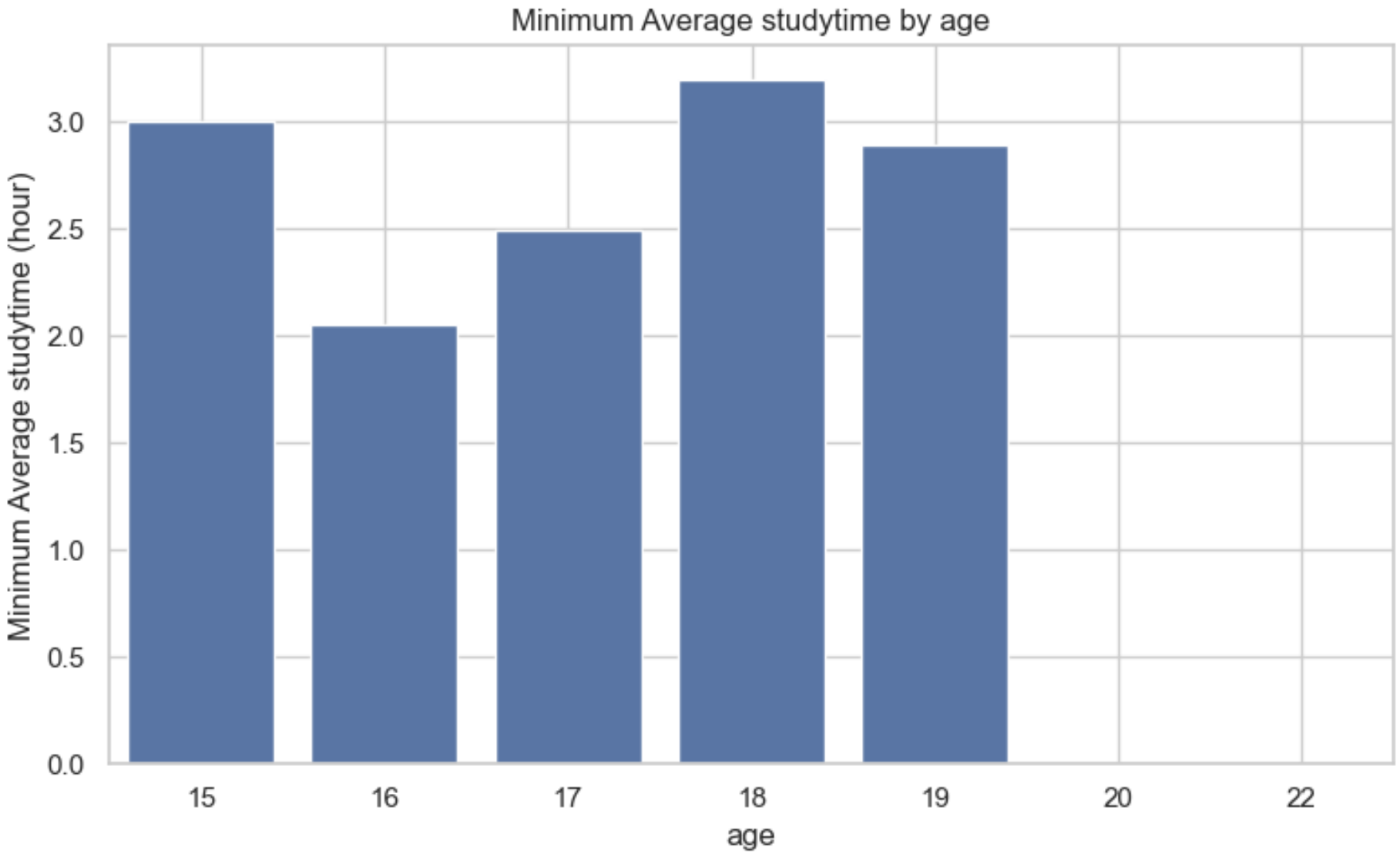
sex, age, internet, romantic.

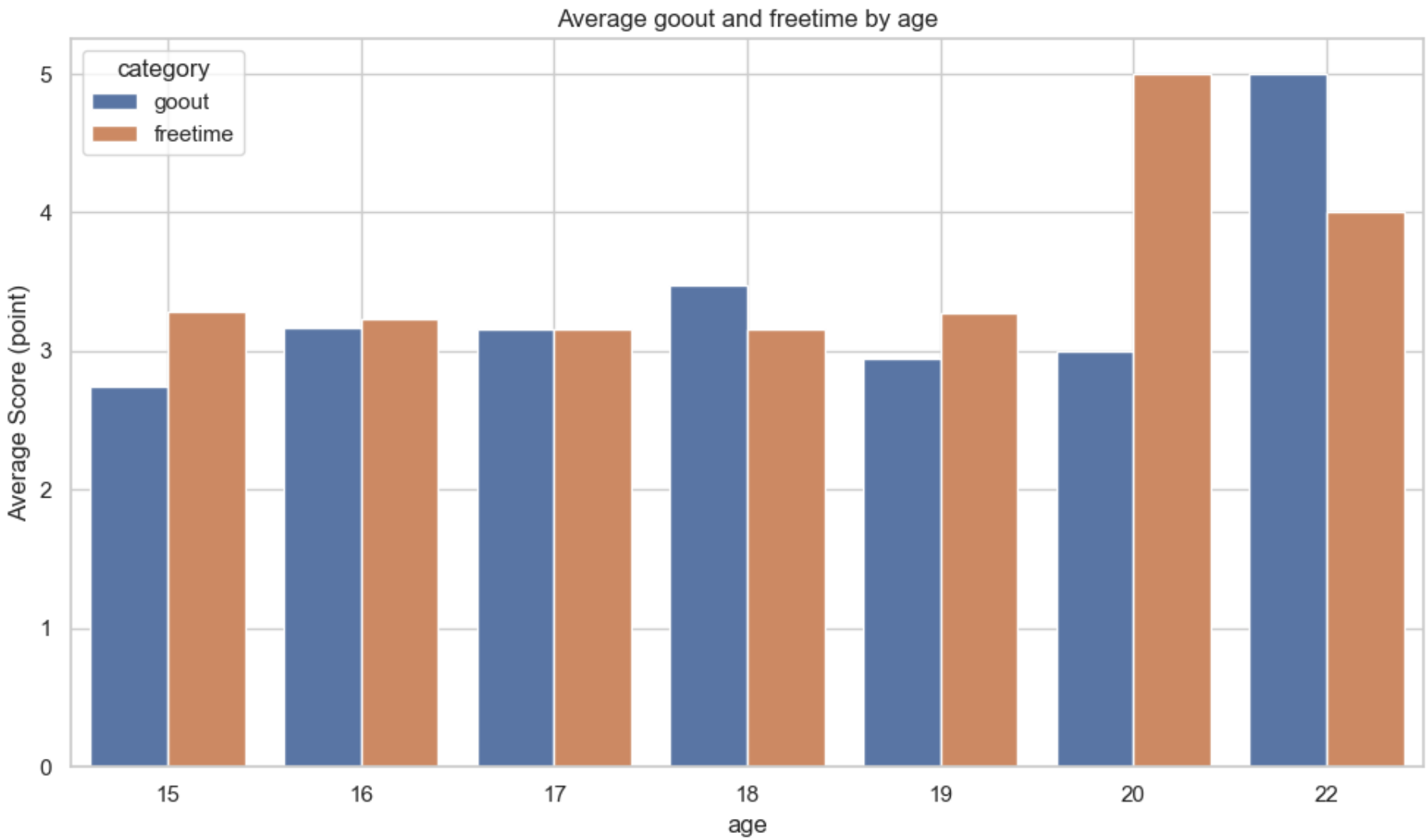
Data Visualization

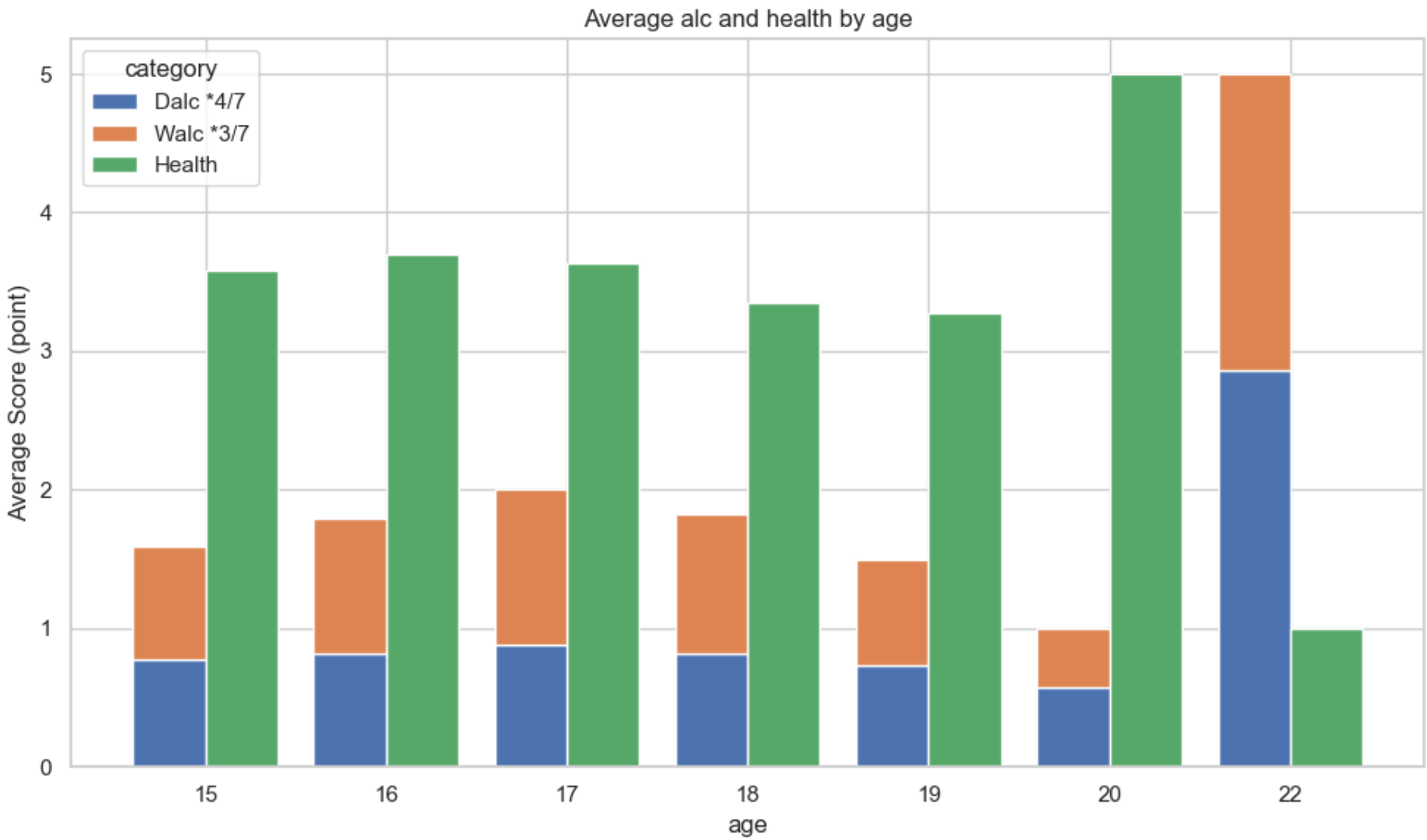


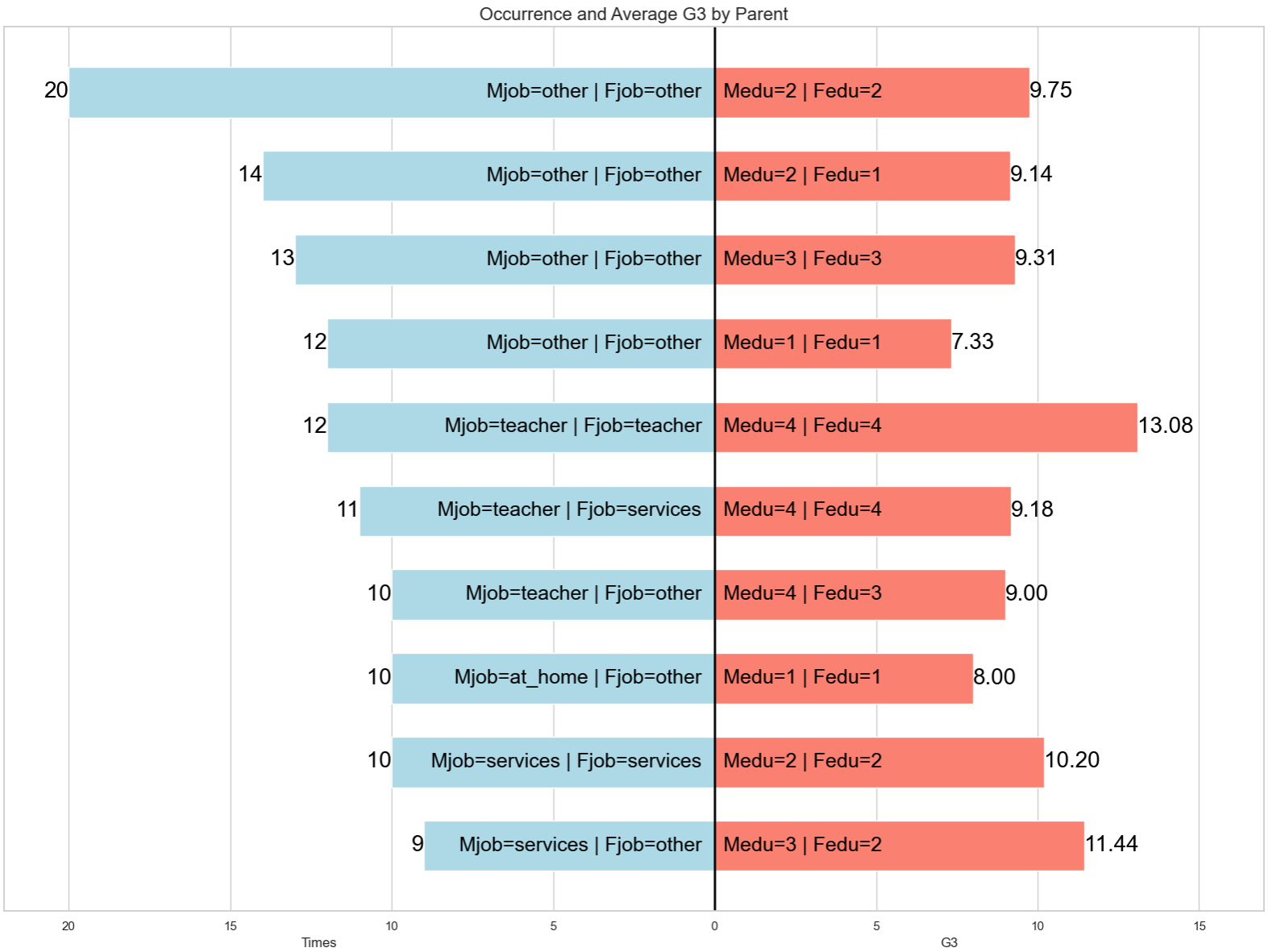
GP (Gabriel Pereira) **MS (Mousinho da Silveira)**

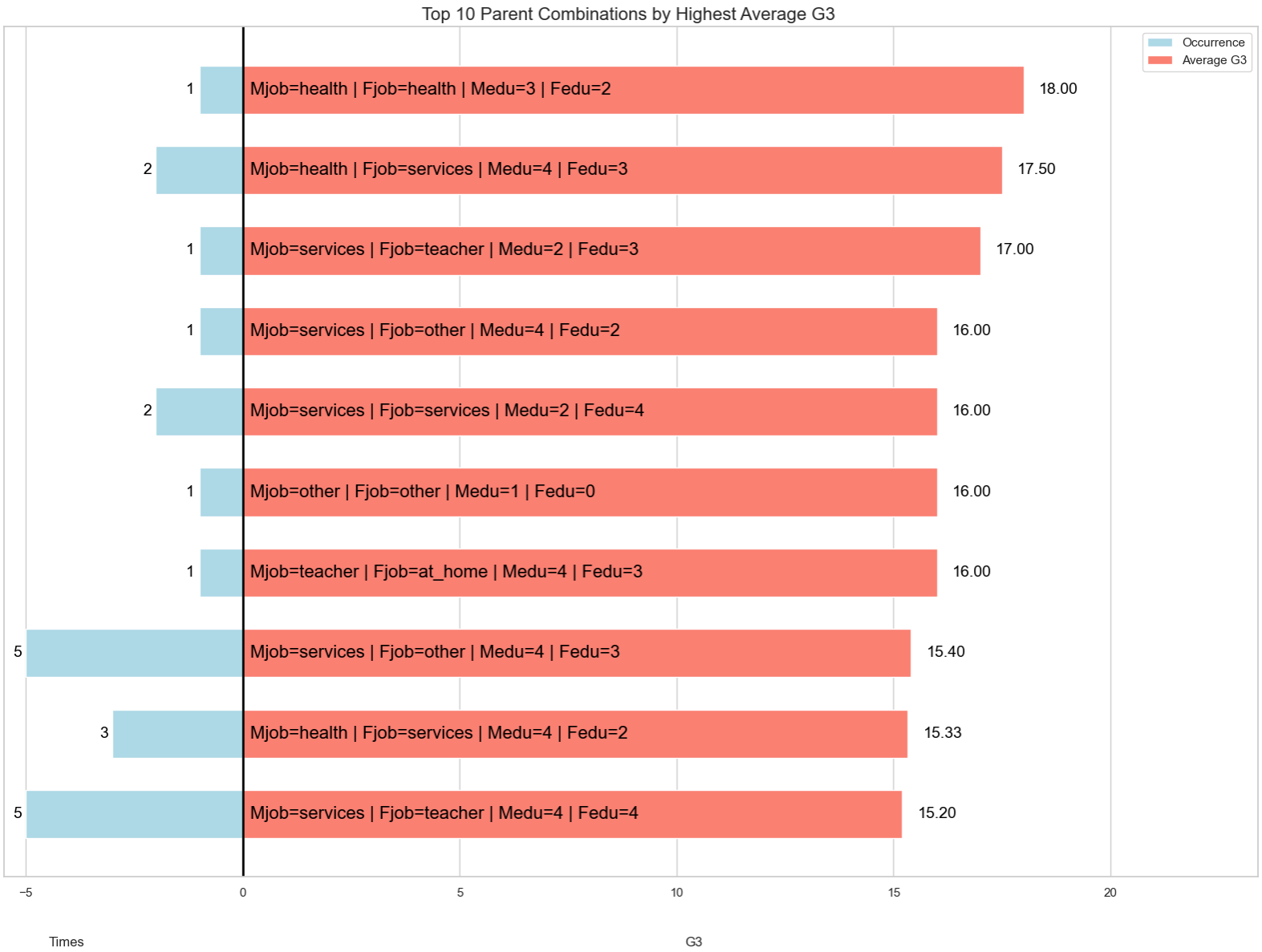

age
100
80
60
40
20
0
15
16
17
18
19
20
21
22
age


Minimum Average studytime by age
Minimum Average studytime (hour)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15
16
17
18
19
20
22
age

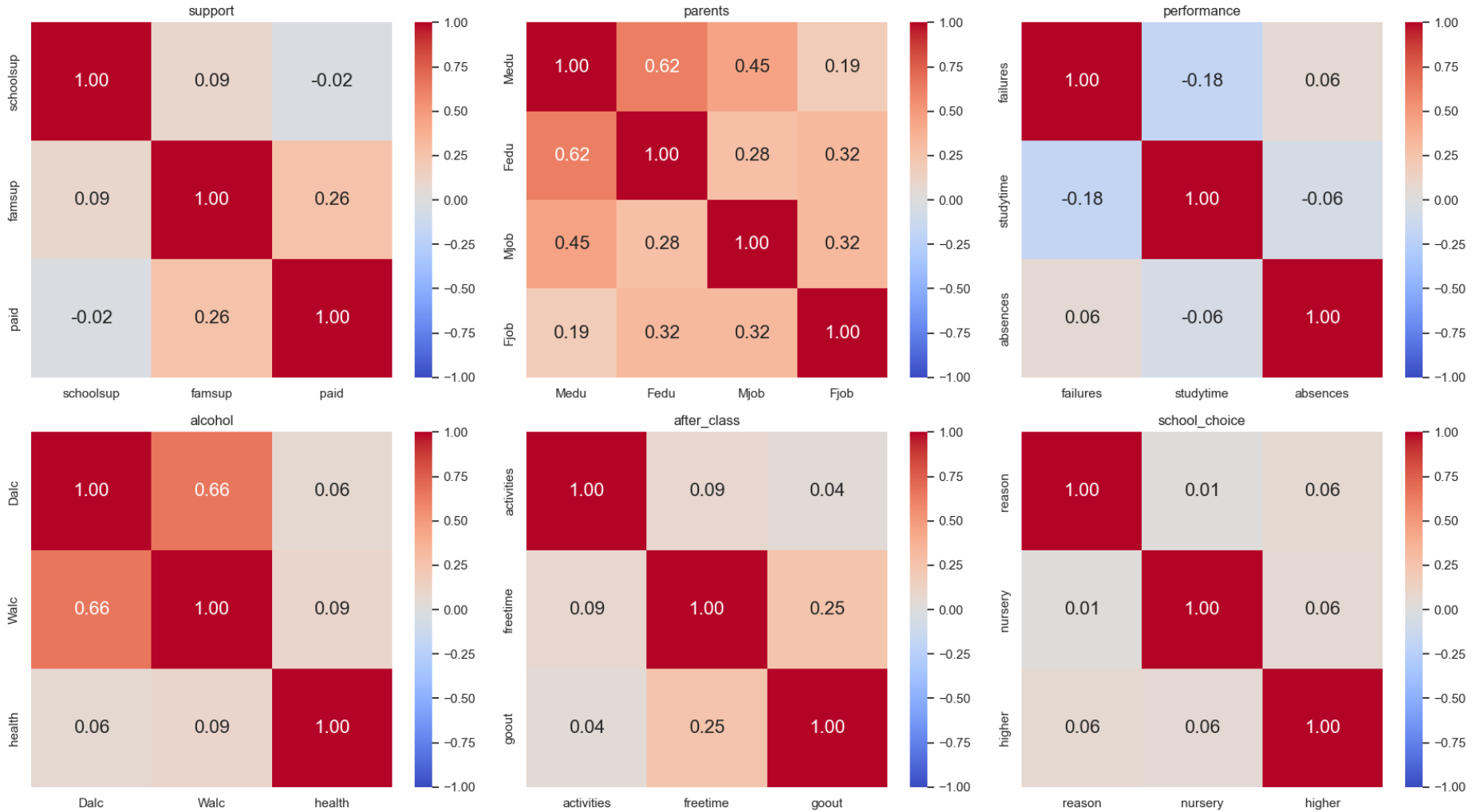

Average goout and freetime by age
category
goout
freetime
Average Score (point)
0
1
2
3
4
5
15
16
17
18
19
20
22
age


Average alc and health by age
category
Dalc *4/7
Walc *3/7
Health
Average Score (point)
0
1
2
3
4
5
15
16
17
18
19
20
22
age

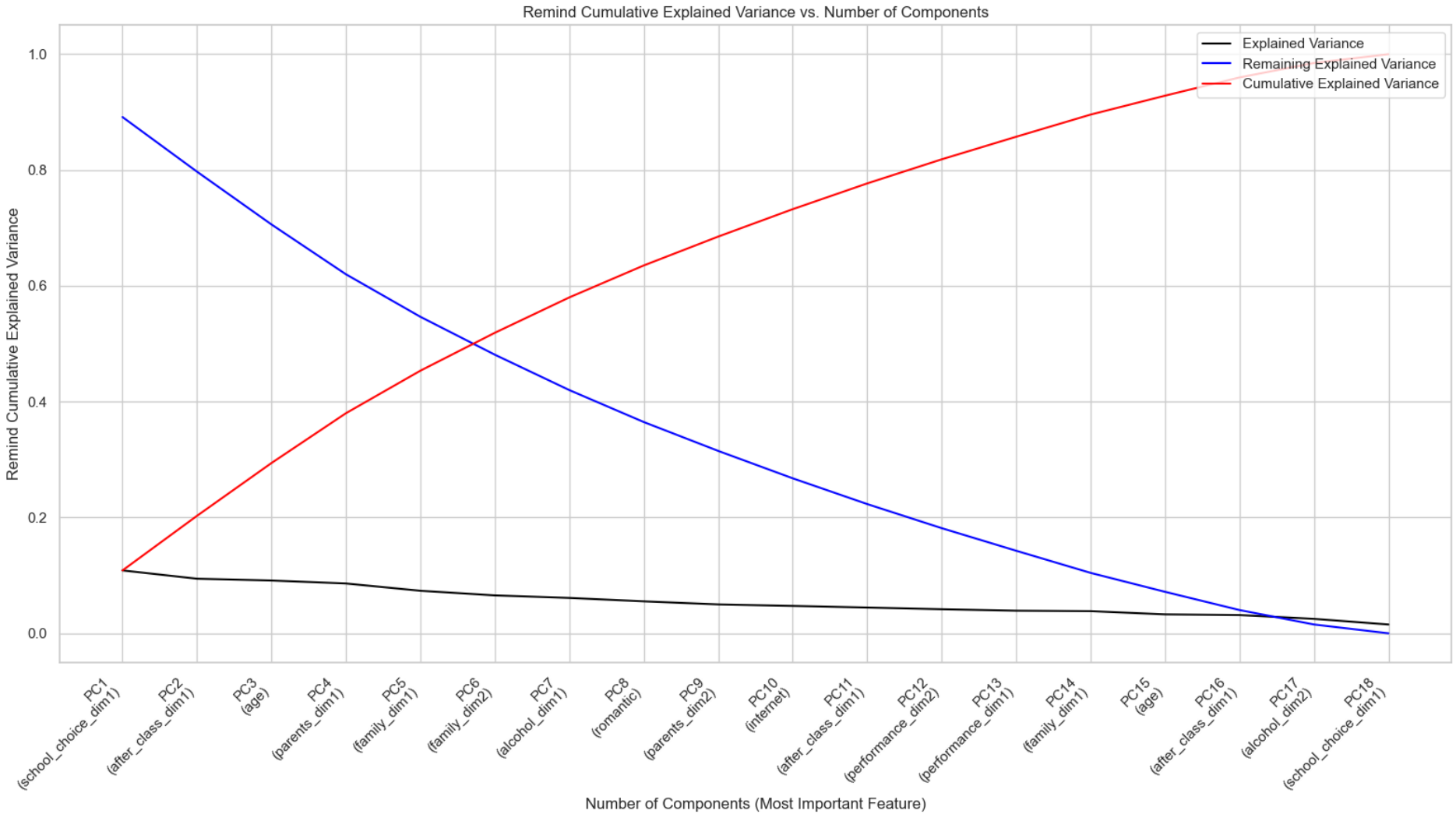

Occurrence and Average G3 by Parent
20 Mjob=other | Fjob=other Medu=2 | Fedu=2 9.75
14 Mjob=other | Fjob=other Medu=2 | Fedu=1 9.14
13 Mjob=other | Fjob=other Medu=3 | Fedu=3 9.31
12 Mjob=other | Fjob=other Medu=1 | Fedu=1 7.33
12 Mjob=teacher | Fjob=teacher Medu=4 | Fedu=4 13.08
11 Mjob=teacher | Fjob=services Medu=4 | Fedu=4 9.18
10 Mjob=teacher | Fjob=other Medu=4 | Fedu=3 9.00
10 Mjob=at_home | Fjob=other Medu=1 | Fedu=1 8.00
10 Mjob=services | Fjob=services Medu=2 | Fedu=2 10.20
9 Mjob=services | Fjob=other Medu=3 | Fedu=2 11.44
20 15 10 5 0 5 10 15
Times
G3


Top 10 Parent Combinations by Highest Average G3
Occurrence
Average G3
1
Mjob=health | Fjob=health | Medu=3 | Fedu=2
18.00
2
Mjob=health | Fjob=services | Medu=4 | Fedu=3
17.50
1
Mjob=services | Fjob=teacher | Medu=2 | Fedu=3
17.00
1
Mjob=services | Fjob=other | Medu=4 | Fedu=2
16.00
2
Mjob=services | Fjob=services | Medu=2 | Fedu=4
16.00
1
Mjob=other | Fjob=other | Medu=1 | Fedu=0
16.00
1
Mjob=teacher | Fjob=at_home | Medu=4 | Fedu=3
16.00
5
Mjob=services | Fjob=other | Medu=4 | Fedu=3
15.40
3
Mjob=health | Fjob=services | Medu=4 | Fedu=2
15.33
5
Mjob=services | Fjob=teacher | Medu=4 | Fedu=4
15.20
−5
0
5
10
15
20
Times
G3

Multidimensional Scaling



PCA


Remind Cumulative Explained Variance vs. Number of Components
Explained Variance
Remaining Explained Variance
Cumulative Explained Variance
Remind Cumulative Explained Variance
1.0
0.8
0.6
0.4
0.2
0.0
PC1 (school_choice_dim1)
PC2 (after_class_dim1)
PC3 (age)
PC4 (parents_dim1)
PC5 (family_dim1)
PC6 (family_dim2)
PC7 (alcohol_dim1)
PC8 (romantic)
PC9 (parents_dim2)
PC10 (internet)
PC11 (after_class_dim1)
PC12 (performance_dim2)
PC13 (performance_dim1)
PC14 (family_dim1)
PC15 (age)
PC16 (after_class_dim1)
PC17 (alcohol_dim2)
PC18 (school_choice_dim1)
Number of Components (Most Important Feature)

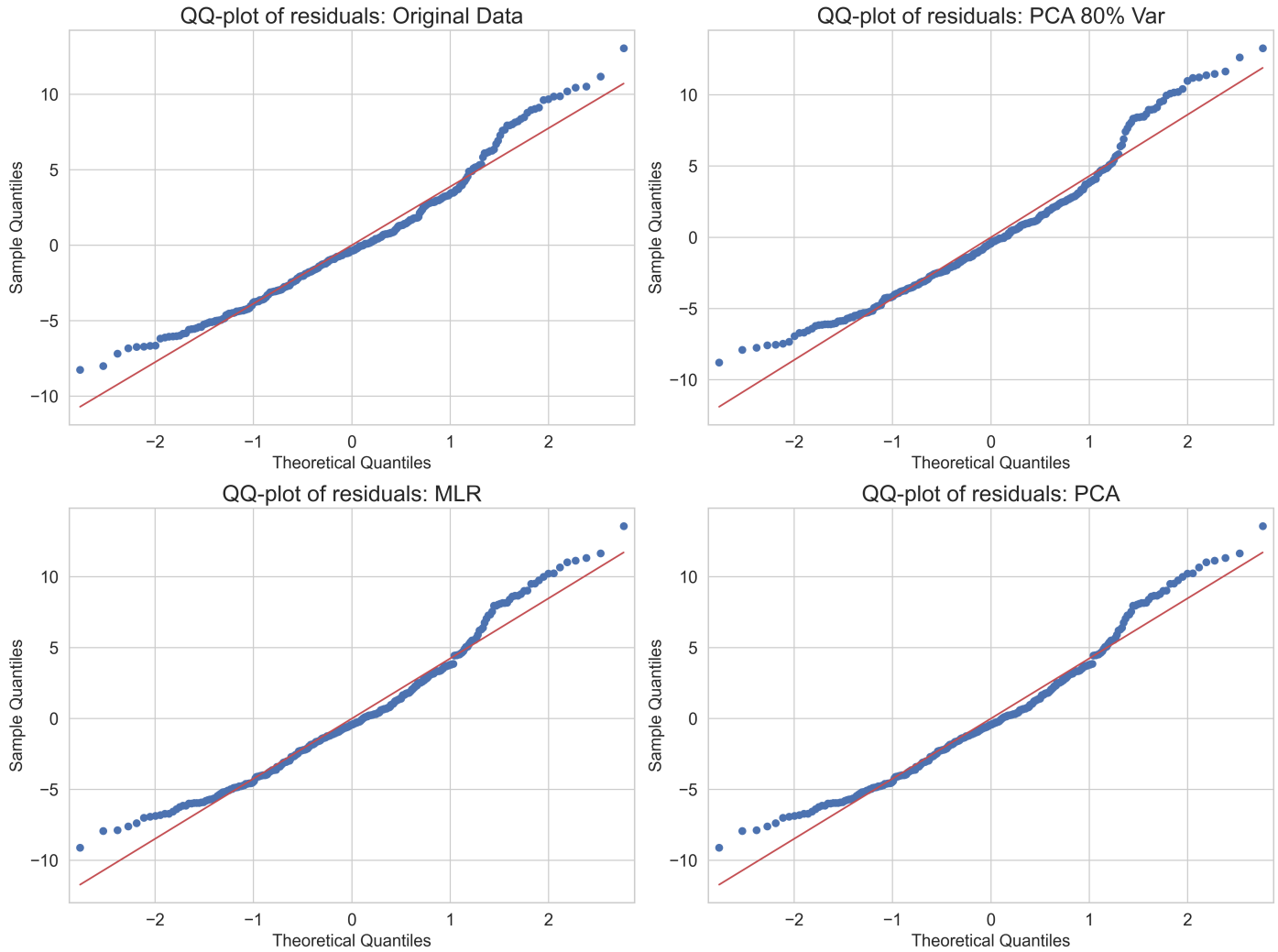
Variables of MLR

We compared four models:

- Original + MLR
- MDS + MLR
- MDS + PCA + MLR
- MDS + PCA 80% + MLR

The variables in the dataset transformed by MDS are as follows:

Q-Q plots of MLR

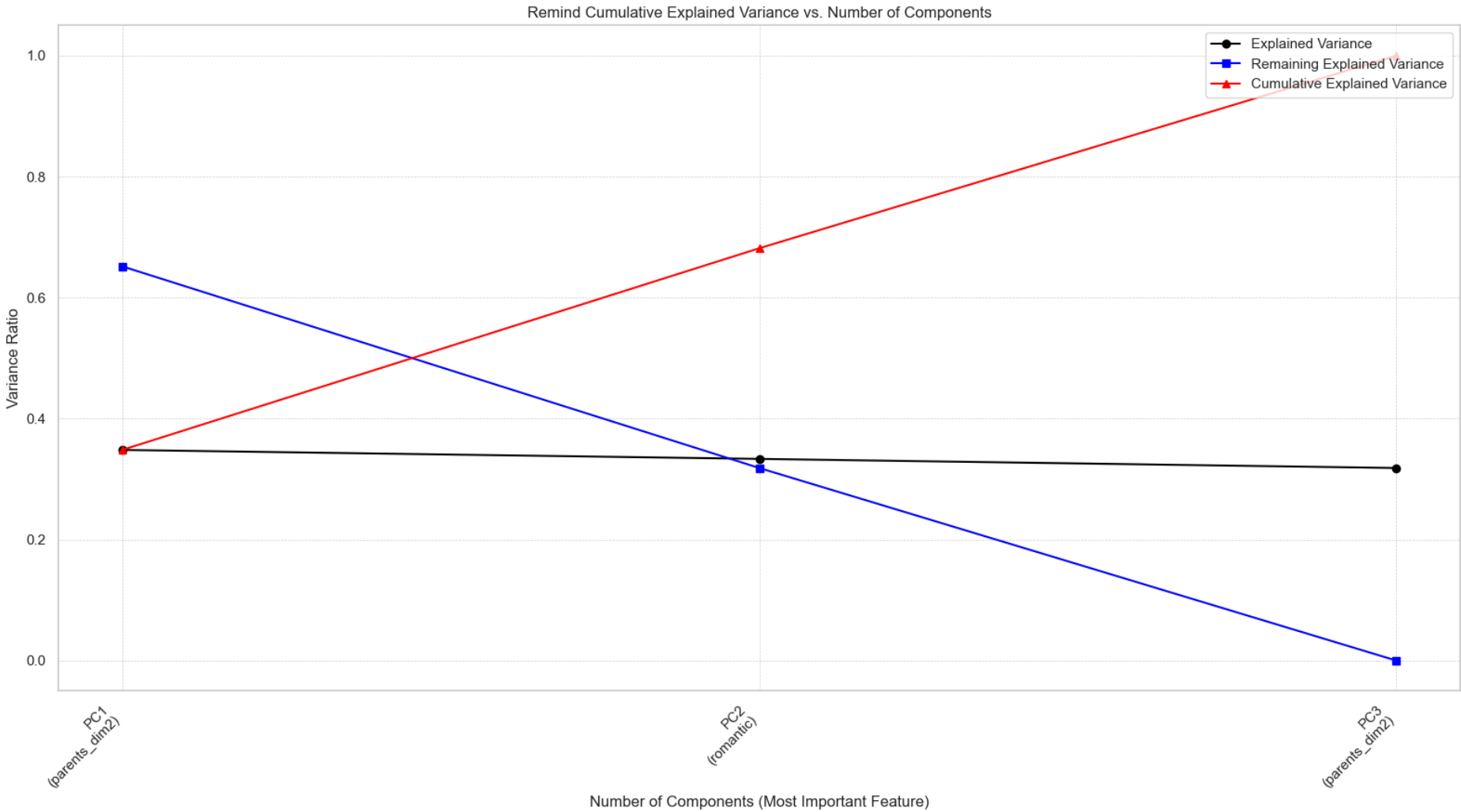


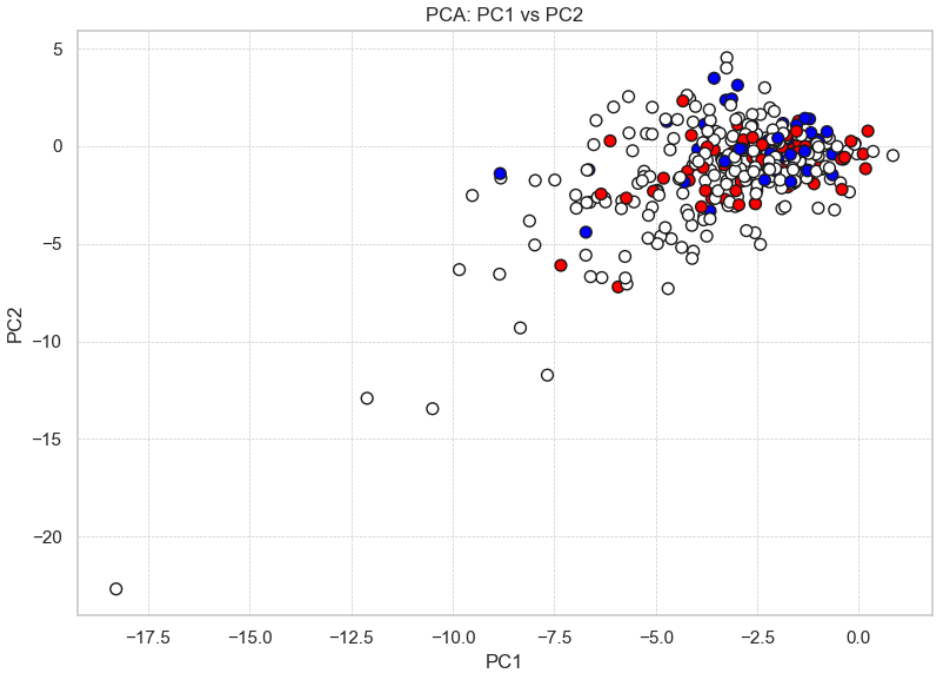
Variables of Reduced MLR

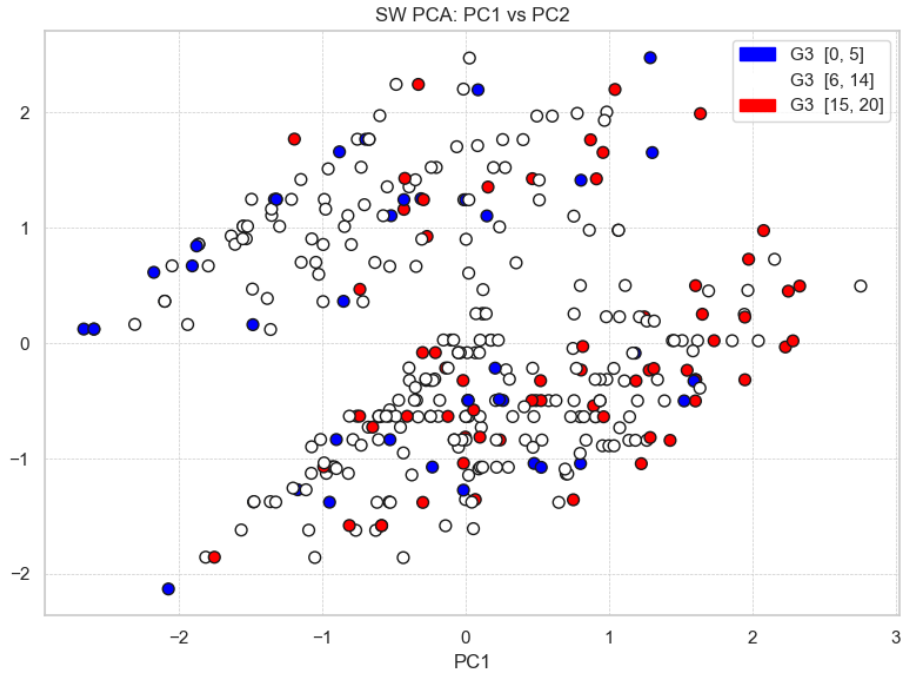
Method	**Selected Features**
Original (Reduced) + MLR	`failures`, `goout`, `Mjob_at_home`, `romantic`, `schoolsup`
MDS (Reduced) + PCA + MLR	`after_class_dim2`, `romantic`, `parents_dim2`

Table 1: Selected Features by Different Methods

Original data + MDS + Stepwise + PCA







PC1 (school_choice_dim1)
PC2 (after_class_dim1)

PC1 (parent_dim2)
PC2 (romantic)

Conclusion

Model	R^2	Adj. R^2	MSE	P-value	A.I.C
Original + MLR	**0.30**	**0.20**	**58.03**	0.9985	2013.87
MDS + MLR	0.16	0.11	65.08	0.5948	2036.70
MDS + PCA + MLR	0.16	0.11	65.08	0.9639	2036.70
MDS + PCA 80% + MLR	0.13	0.10	81.21	0.9227	2035.49
Original (Reduced) + MLR	0.19	0.18	288.71	**0.0375**	**1995.23**
MDS (Reduced) + PCA + MLR	0.08	0.07	204.60	0.7878	2036.42
Paper Proposed + MLR	0.19	0.17	176.54	0.6199	2003.04

Table 2: Comparison of model performance metrics

The variables selected by the paper:
absences , schoolsup , higher , failures , Mjob

Conclusion

Model	TOP1	TOP2	TOP3
Original + MLR	**higher**	**Fjob_teacher**	**failures**
MDS + MLR	romantic	support_dim2	after_class_dim2
MDS + PCA + MLR	alcohol_dim1	after_class_dim1	age
MDS + PCA 80% + MLR	performance_dim2	alcohol_dim1	age
Original (Reduced) + MLR	**failures**	**Mjob_at_home**	**schoolsup**
MDS (Reduced) + PCA + MLR	parents_dim2	romantic	parents_dim2
Paper Proposed + MLR	**higher**	**failures**	**Mjob_health**

Table 3: Most important feature to each model

Conclusion

The variables we selected are:

failures, Mjob, schoolsup

References

- Cortez, P. (2008). *Student Performance* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T
- Cortez, P., & Silva, A. M. (2008). Using data mining to predict secondary school student performance.
- Shamim, A. (2025). *Math students performance data*. Kaggle. https://www.kaggle.com/datasets/adilshamim8/math-students

Thanks for listening!